MATH 343 LEC 14 Conjugate Bayesian Estimation

$\vec{\varepsilon} \sim N_n(\vec{0}, \sigma^2 I_n)$, $\vec{Y} = X\vec{\beta} + \vec{\varepsilon}$, $\vec{\beta} \sim N_{p+1}(\vec{0}_{p+1}, \tau^2 I_{p+1})$, $f(\sigma^2) \propto \frac{1}{\sigma^2}$

$\vec{b}_{Ridge} := \underbrace{\vec{\beta}^{MAP}}_{\text{Ridge Estimate}} = (X^T X + \lambda I_{p+1})^{-1} X^T \vec{y}$ where $\lambda = \frac{\sigma^2}{\tau^2} > 0$

If we employ this estimator for prediction purposes, we ignore the fact that σ^2 is estimated and treat λ as a hyperparameter.

Slight practical problem: we shrink each slope estimate equally as σ^2 is the same prior variance. If the $\vec{x}_j$'s are on different scales, it is doing unequal shrinking. So, as a prestep, normalize all variables:

Let X be the design matrix where $\vec{x}_{*j} = \frac{\vec{x}_{*j} - \bar{x}_{*j}}{s_{\vec{x}_{*j}}}$ for $j \geq 1$ (omit intercept)

this is called "standardization". The standardized feature measurements have avg zero, std. err one. They all shrink equally. Demo

So assuming standardized covariates...

Let's try another prior on $\vec{\beta}$:

$\beta_0, \beta_1, \ldots, \beta_p \overset{iid}{\sim} \text{Laplace}(0, \tau^2) := \frac{1}{2\tau^2} e^{-\frac{|\beta|}{\tau^2}}$

$f(\sigma^2) \propto \frac{1}{\sigma^2}$

$f(\vec{\beta}, \sigma^2 | X, \vec{y}) \propto f(\vec{y} | \vec{\beta}, \sigma^2, X) f(\vec{\beta}, \sigma^2 | X)$

$= \frac{1}{(2\pi)^{n/2} \sqrt{\det[\sigma^2 I_n]}} e^{-\frac{1}{2}(\vec{y} - X\vec{\beta})^T (\sigma^2 I_n)^{-1} (\vec{y} - X\vec{\beta})}$

$\prod_{j=0}^{p} \frac{1}{2\tau^2} e^{-\frac{|\beta_j|}{\tau^2}} \cdot \frac{1}{\sigma^2}$

$\propto (\sigma^2)^{-\frac{n}{2}-1} e^{-\frac{1}{2\sigma^2} \|\vec{y} - X\vec{\beta}\|^2} e^{-\frac{1}{\tau^2} \sum_{j=0}^{p} |\beta_j|}$

Now find $\vec{\beta}^{MAP}$

$\vec{\beta}^{MAP} = \arg\max \{ f(\vec{\beta}, \sigma^2 | X, \vec{y}) \} = \arg\max \{ \ln f(\vec{\beta}, \sigma^2 | X, \vec{y}) \}$

$= \arg\max \left\{ \underbrace{(-\tfrac{n}{2} - 1) \ln \sigma^2}_{\text{constant}} - \frac{1}{2\sigma^2} \|\vec{y} - X\vec{\beta}\|^2 - \frac{1}{\sigma^2} \sum_{j=0}^{p} |\beta_j| \right\}$

$= \arg\max \left\{ \underbrace{-\frac{1}{2\sigma^2}}_{\text{constant}} \left(\|\vec{y} - X\vec{\beta}\|^2 + \frac{2\sigma^2}{\tau^2} \sum_{j=0}^{p} |\beta_j| \right) \right\}$

$\uparrow$ maximize

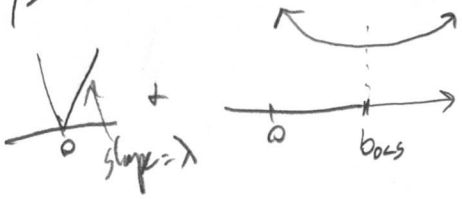
$= \arg\min \left\{ + \|\vec{y} - X\vec{\beta}\|^2 + \frac{2\sigma^2}{\tau^2} \sum_{j=0}^{p} |\beta_j| \right\}$ Now plug the 3rd line on...

let $\lambda = \frac{2\sigma^2}{\tau^2}$

$\vec{b}_{LASSO} := \arg\min \left\{ SSE + \lambda \sum_{j=0}^{p} |\beta_j| \right\}$

No closed form solution exists.
Need to use optimizer

As $\lambda \to 0 \Rightarrow \vec{b}_{Lasso} \to \vec{b}$. As $\lambda \to \infty \Rightarrow \vec{b}_{Lasso} \to \vec{0}_{p+1}$

Also. If $\vec{\beta}$ was one dimensional, this is what the log prior looks like 

(demo)

If λ is large and/or if $b_{OLS} \neq 0$, the sum will have min. near zero (demo)

It's difficult for the $\|y - X\beta\|^2$ term to "beat" the $\lambda \sum_{j=1}^{p} |\beta_j|$ term. Hence, Lasso has an argmin at $\beta = 0$ due to the sharp Laplace prior which has its sharp mode at zero.

$\Rightarrow$ Lasso has an ability to do "variable selection". I.e. it (first) pulls variables when there is a large set that should be pared down as you know you'll overfit if you use all of them